doi: 10.1255/nirn.1511

Summary of the 2014 IDRC software shoot-out

Benoit Igne,[a] Andrey Bogomolov,[b,c] Dongsheng Bu,[d] Pierre Dardenne,[e] Vladislav Galyanin[c] and Peter Tillmann[f]
[a]GlaxoSmithKline, King of Prussia, PA, USA
[b]Consultant, Global Modelling, Aalen, Germany
[c]Samara State Technical University, Samara, Russia
[d]Bristol-Myers Squibb, New Brunswick, NJ, USA
[e]Walloon Agricultural Research Centre, Gembloux, Belgium
[f]VDLUFA Qualitätssicherung NIRS GmbH, Kassel, Germany

The Software Shoot-Out has been a staple of the International Diffuse Reflectance Conference (IDRC), a biennial meeting taking place in Chambersburg, Pennsylvania, USA. It is a competition amongst participants of the conference that aims to acknowledge and reward the person who develops the best model(s) and obtains the lowest prediction error for a particular diffuse reflectance dataset. Every IDRC, a new challenge is proposed. The conference's website (http://www.idrc-chambersburg.org) provides access to this dataset as well as those used for previous challenges and previous *NIR news* articles have reported results from the past two competitions.[1,2]

Two competitions took place during the 2014 conference: as usual, a dataset was made available for download and completion at home while, for the first time, an additional, on-site competition was proposed to all conferees. This on-site shoot-out was carried out as an anonymous challenge in which students and professionals used their chemometric skills to come up with the best prediction models for two parameters pertaining to a single dataset. The top three students and the top three professionals were recognised during the conference banquet. One third of the conferees participated, a very high and encouraging figure.

The more traditional shoot-out presentation took place following the on-site challenge and was a great occasion at which to learn from and interact with experienced chemometricians presenting their approach to a common multivariate analysis problem. For the first time, a petrochemical dataset was used. The conference would like to thank Halliburton, Christopher M. Jones and David Perkins for providing the data and Michael Myrick for facilitating the process. However, given the nature of the data and the competitive nature of the research field, some restrictions were necessary to reduce the potential for competitors to use the data for their own commercial advantage. Specifically, the wavelength scales were unspecified (although it covered the NIR), the calibration values were normalised and the nature of the parameters being predicted was not communicated. Two datasets were provided. Only one parameter was available per dataset.

The challenge consisted of developing the best model for the parameters and datasets provided using the calibration data. Because of the limited amount of information available, success in the shoot-out depended on the participants' ability to build models by relying only on their chemometric skills and not their *a priori* knowledge of the nature of the data. However, the most important task was to build models that would be robust to the variability present in the validation set and possibly not present in calibration. In addition, the quality of the presentation of the results and the reasoning behind the approach taken were used to determine the winners. Participants were therefore required to:

1) develop the best possible models for the parameters using the calibration set,
2) test their models on a test set (reference values provided),
3) predict validation sets (reference values not provided) and
4) detail their reasoning when selecting pre-treatment methods, regression method and number of latent variables during a presentation.

Datasets

Dataset 1

These samples corresponded to oils from petroleum reservoirs around the world. These particular spectra were collected in transmittance in the laboratory under various conditions of pressure and temperature, using a high pressure flow cell. It should be noted that, as pressure and temperature change, so does the effective path length and the density of the fluid. The nominal path length was 1 mm. The temperature and pressure at which the spectra were acquired were provided and the path length can be calculated using the following equation (note that in the dataset, the temperatures are given in °C and pressures in psi):

$$d\,(\text{mm}) = 0.8801 + 0.000402065 \times T\,(\text{Kelvin}) + 0.00060493 \times P\,(\text{MPa})$$

Data correspond to NIR and IR data and the axes, while not labeled, are in cm^{-1}, linearly spaced (Figure 1). Finally, while most spectra represent mixtures, some pure components are provided but not identified as such. No pure components are present in the test and validation sets.

Dataset 2

These samples corresponded to gas mixtures in the gas phase measured in transmittance. As was the case for Dataset 1, samples were collected in the laboratory under various conditions of pressure and temperature but these details were not provided. Data corresponded to NIR and IR data and the axes, in linearly-spaced nm, were not provided.

Note that this dataset was also used for the Chimiométrie conference that took place in Geneva (Switzerland) on 19–21 January 2015. More information about this meeting is available at the following web address (http://chimio2015.sciencesconf.org/resource/page/id/21).

The approaches taken by four of the five participants are presented below.

Participant 1

A calculation-intensive approach to model optimisation was selected based on Joint Variable Selection and Pre-processing Optimisation (JVSPO).[3] This methodology is based on the following idea: pre-processing algorithms and variable selection are powerful tools for performance improvement of

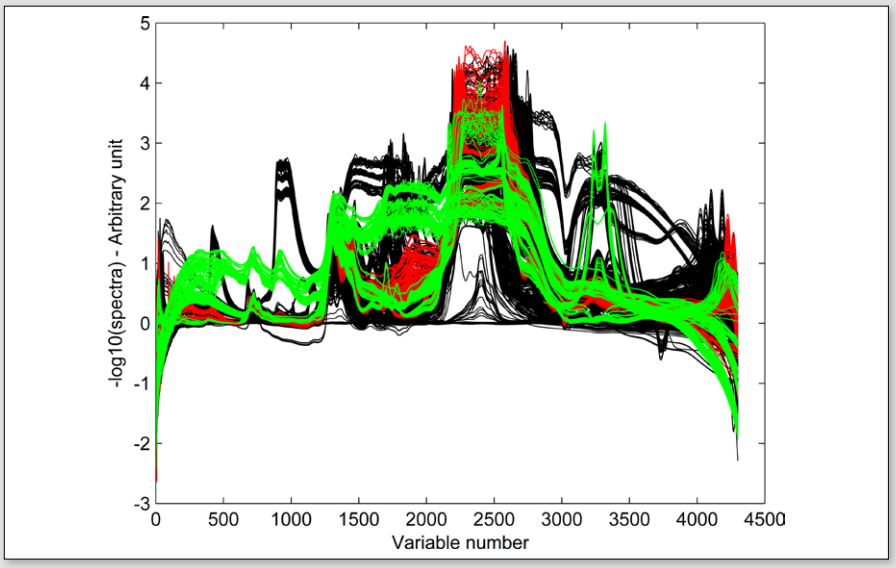

Figure 1. Overlay of the calibration (black), test (red), and validation (green) data for Dataset 1. (Note that not all available variables were displayed.)

calibration models built on spectroscopic data. Typically, a pre-processing is first optimised and applied followed by variable selection. In the JVSPO approach, the pre-processing method and its parameters as well as an optimal variable set are found at a time in the same optimisation routine, e.g. using a Genetic Algorithm (GA). The merit function used at the optimisation is custom and can be composed of any statistical criteria, such as root mean-square errors of calibration, validation and prediction or respective r^2 values.

The variables can be selected individually or as intervals of different widths to take into account possible correlations between adjacent spectral variables. The variables within an interval can be optionally averaged. This approach can be thought of as a generalisation and extension of the well-known interval PLS (iPLS) algorithm. Model computation times were about 2 h and 1 h for datasets 1 and 2, respectively.

Despite the trial-based approach, exploratory data analysis was always performed prior to the calibration, because understanding of the data structure is critical for modelling success. Investigation of spectra in Dataset 1 revealed two different regions, which might be a result of augmentation of different data sources. One part of the data exhibited intensities and shapes characteristic of transmission spectra (variables 1–4150). Data linearisation through absorbance transformation was tested along with using the raw data. Additionally, the path length correction was used. Neither of those transformations was justified for the second spectral region (variables above 4150). Considering dramatic intensity differences between the parts, auto-scaling was performed prior to modelling. All combinations of the above transforms were tested followed by the whole model optimisation cycle. The best results were obtained for raw (transmittance) data of the first part after correction for path length. The best model was built with 30 intervals of 7 averaged variables each; pre-processing involved Savitzky–Golay 1st derivative (polynomial order = 1, window = 7 points).

Principal component analysis (PCA) exploration of Dataset 2 revealed that the validation spectra were completely different from the rest of the data. Only one sample presenting the variability of the validation set was present in the test set. In fact, the validation set provided was not representative of the calibration set. To overcome this inconsistency, the reference *y*-value distribution of the calibration set was integrated into the merit function along with *RMSE* of calibration and test. This approach ensured that predicted *y*-values of the test set were reasonably similar, which was a natural assumption. The best model in Dataset 2 was built with 20 intervals of 21 variables each (no averaging); pre-processing involved Savitzky–Golay 1st derivative (polynomial order = 1, window =18 points).

Participant 2

Two regression models were developed for Dataset 1: principal component regression (PCR) was used to model spectral data with associated reference values ranging from 0.0 to 0.3 and a partial least squares (PLS) model was used with samples and reference values above 0.3. After a visual inspection of the spectral data, three regions were selected for modelling (1160–1640, 2610–3000 and 3200–3900 variable points), corresponding to regions with transmittance values below 100 (arbitrary unit). Reference values were corrected for path length differences by the equation provided with the sample measurement temperature and pressure. None of the test samples were used in calibration; however, they were utilised for determining model suitability limits. Hotelling's T^2 and Q-residual were employed for prediction suitability testing in order to assign a model (PCR or PLS) to the unknown samples in the validation set.

For the PCR model, 13 principal components were used to fit a total of 1306 calibration samples, which had *y* reference values ranging from 0.0 to 0.3. Some samples presenting a reference value of zero were excluded to avoid overweighting the model. Spectral data were pre-treated by transmittance to absorbance conversion followed by baseline offset correction. A suitability limit for *y*-deviation in The Unscrambler® was determined as 0.03 from the test dataset.

For the PLS model, 9 factors were used to fit a total of 124 calibration samples with *y* reference values of 0.3 and above. Spectral data were pre-treated as follows: transmittance to absorbance conversion, baseline offset correction, unit-vector normalisation and Savitzky–Golay 2nd derivative (21 points, 3rd order polynomial smoothing). Nine wavelengths were selected by forward stepwise regression in MATLAB. During prediction, the PCR model was applied first. For each sample of the validation set, if the *y*-deviation was less than 0.03, the predicted value was recorded otherwise, the PLS model was applied instead. Predicted values were scaled back by their individual measurement temperature and pressure.

For Dataset 2, a 7-factor PLS model was built from 130 calibration samples and spectral data were pre-treated by transmission to absorbance conversion and Savitzky–Golay 2nd derivative (21 points, 3rd order polynomial smoothing). Three spectral

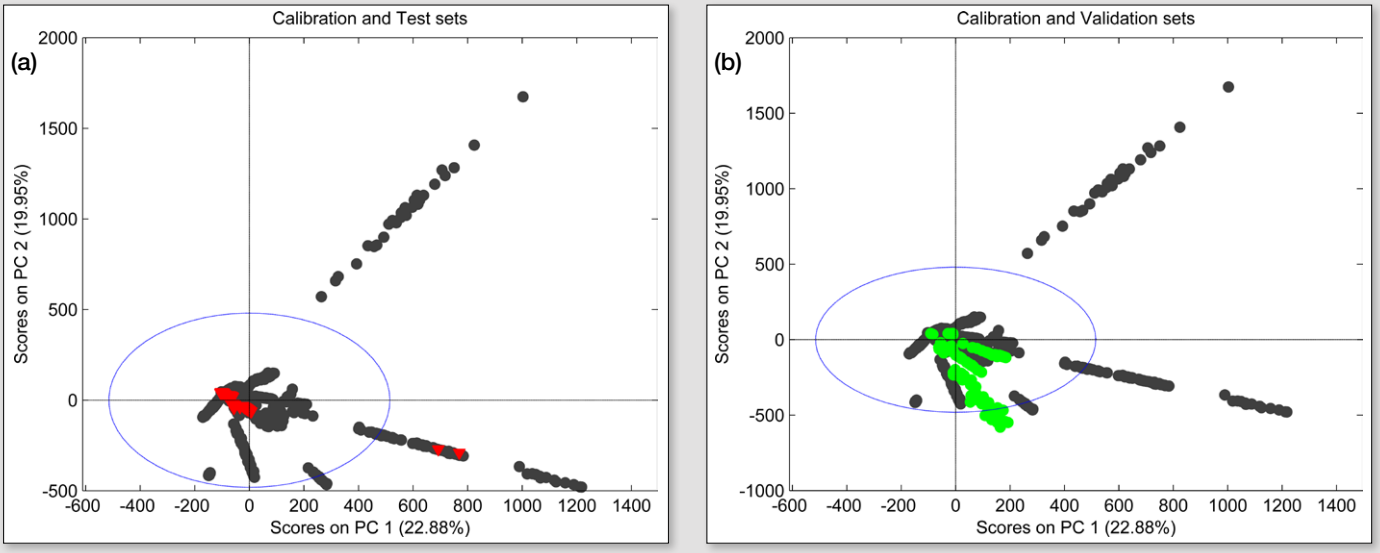

Figure 2. Dataset 1 PCA plot of the calibration and test samples (a) and validation samples (b).

regions (101–1651, 1851–2201 and 2511–3101) were selected for modelling.

Participant 3

The first step of the data exploration consisted in looking at the available reference data. The *y* vector presented a very dissymmetric distribution and temperature and pressure were provided. 2D and 3D plots with the reference values, pressure and temperature did not show any correlation but rather some type of orthogonal experimental design. Before looking at the spectra, the transmittance data were corrected with the pathlength changes arising from temperature and pressure according to the equation provided. The corrections varied from 0% to 16%. As more absorption occurs with larger pathlengths, transmittance values would have been higher with shorter path lengths; therefore, this correction increased the transmittance values. The corrected spectra were then converted to absorbance.

A global plot of the spectra was unreadable due to the variability in terms of peak positions and absorbance values (Figure 1). A PCA on the calibration set showed well-separated groups of spectra (Figure 2a) and the groups seemed to be sequentially merged in the common data set. A projection of the test set indicated three main groups of spectra (1–99, 100–162, 163–269).

For each group, several multiple linear regression (MLR) step-up models were tested with the most common region from variables between 580 and 2000. Models obtained involved between five and seven predictors with only a short smoothing of three data points as pre-processing. For each group, the final model retained was the one presenting the smallest Mahalanobis distances. The prediction errors (*RMSEP*s) were 0.002, 0.006 and 0.102, respectively, for the three sub-groups of the test set with a global r^2 of 0.996.

The validation set had a structure again different from the calibration and test sets. A PCA revealed five groups of rather different spectral features from the calibration set (Figure 2b). So the models selected for the test set were not going to be useful for predicting the validation set. Each validation group was treated individually. For groups 1, 3, 4 and 5, a local approach was used (Foss WinISI 4 – Local calibration) with the following respective ranges, the number of selected spectra, the number of PLS terms and the pre-processing: G1: 650–1900, 50 spectra, 5 factors, 1st gap derivative, 11 point window; G3: 2150–3100, 30 spectra, 4 factors, 1st gap derivative, 11 point window; G4 and G5: 700–1380, 50 samples, 5 factors, 1st gap derivative, 5 point window. For G2, the spectra were too far apart to find any common spectral region so the Local methodology was unusable. Models for that group were then developed by MLR step-up and by trial and error to select a model with the minimum of the Mahalanobis distances. The averaged Mahalanobis distance values for the five groups were, respectively: #1–26, 0.8; #27–73, 2.7; #17–113; 2.1; #114–163, 1.9 and #164–208, 1.2.

The current exercise emphasised the importance of the Mahalanobis distances, confirming once more that empirical models only work if the spectral data to be predicted are close to the variance included in the calibration set.

Participant 4

Dataset 1 exhibited a skewed distribution with many samples presenting values under 0.1, with a few at "fixed levels" 0.3 and above 0.7. After pressure and temperature normalisation of the spectra, a simple PLS model did not work at all. So a selection of the spectral region specific to the analyte of interest was performed. A correlation plot of the reference values versus variables for the calibration and test set was calculated. Only the samples with values above 0.3 with pressures at 6000 psi and temperatures at 150°C were used for this evaluation. The region around data point 1500 was chosen for further investigation. A stepwise MLR was calculated using the data points 1495 and 1252; no scatter correction or derivative was used. The model was biased and slope corrected for the samples with a reference value above 0.1 and used to predict the validation set.

Dataset 2 exhibited significant spectral differences between the calibration and validation sets in the 1500–2000 variables region and above 3000. A local regression model was used with 100 samples chosen and 20–30 factors.

continued on page 14

continued from page 10

Table 1. Validation statistics for sets 1 and 2.

		Participant 1	Participant 2	Participant 3	Participant 4
Set 1	*RMSEP*	0.119	0.567	0.105	0.202
	SEP	0.112	0.523	0.099	0.193
	Bias	–0.039	0.220	–0.035	–0.059
	r^2	0.935	0.002	0.984	0.921
Set 2	*RMSEP*	0.014	0.017	0.017	0.031
	SEP	0.014	0.005	0.007	0.012
	Bias	–0.003	0.016	0.016	0.029
	r^2	0.570	0.968	0.916	0.665

Results

Table 1 presents the validation results for each participant. Root mean square errors (*RMSEP*), standard errors (*SEP*) and bias values are presented along with coefficients of determination.

The participants chose quite different approaches to get prediction results that also varied significantly. With the overall best statistics, participant 3 won the 2014 IDRC Shoot-Out, followed by participant 1 and 2.

The data are available on the IDRC website (http://www.idrc-chambersburg.org). The authors would like to thank the 2014 IDRC chair Dr Rodolfo J. Romañach and the Council for Near-Infrared Spectroscopy for providing funding and support for the conference. The next conference will take place from 30 July to 5 August 2016.

References

1. B. Igne, P. Dardenne, D. Honigs, J.T. Kuenstner, K. Norris, Z. Shi and M. Westerhaus, "The 2010 IDRC software shoot-out at a glance", *NIR news* **21(8),** 14–16 (2010). doi: http://dx.doi.org/10.1255/nirn.1216
2. B. Igne, P. Berzaghi, D. Bu, P. Dardenne, P. Tillmann and M. Westerhaus, "Summary of the 2012 IDRC software shoot-out", *NIR news* **23(7),** 13–15 (2012). doi: http://dx.doi.org/10.1255/nirn.1331
3. V. Galyanin, A. Melenteva and A. Bogomolov, "Selecting optimal wavelength intervals for an optical sensor: A case study of milk fat and total protein analysis in the region 400–1100 nm", *Sensor. Actuator. B*, in press (2015).